Introduction to Science Commons and the Neurocommons

Alan Ruttenberg

Building an Information Framework for Neuroscience

October 19, 2007

What's a (Science) Commons?

- Built on open resources: public domain, open databases, open literature
- Encoded in open architectures and technical standards

Science Commons

- Science Commons is a project of Creative Commons
  - Creative Commons provides free tools that let authors, scientists, artists, and educators easily mark their creative work with the freedoms they want it to carry
  - 140,000,000 objects on the Web under CC licenses in 40+ countries
  - 700+ peer-reviewed journals carry CC licensing, including Public Library of Science
- Science Commons specializes CC to science
  - For consumers of knowledge: make it easy to use and re-use information and increase chances for discovery
  - For providers of knowledge: provide legal certainty and automated attribution and tracking
  - For funders: provide new metrics for tracking return on investment based on re-use

Old Collaboration

- reading the canon on paper
- querying single-access databases
- human as mediator
- artisanal tool manufacturing
- tightly controlled distribution

New Collaboration

- reading the canon with machines
- integrating databases
- computer as mediator
- mashups
- open distribution

What worked at Millennium.

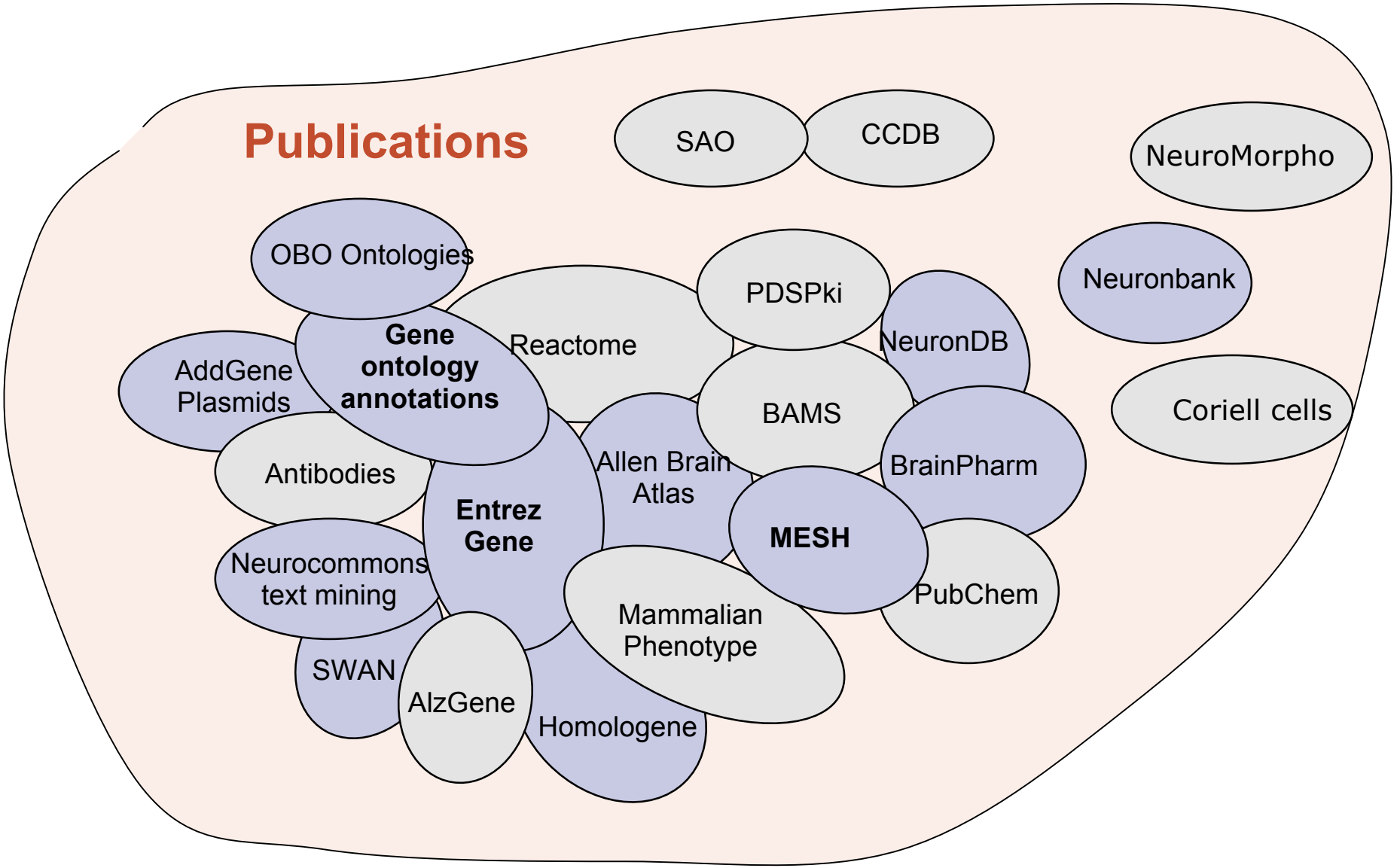
- **Collecting structured knowledge**
  - Integrated public, licensed, and internal KB's
  - The best licensable KB: Ingenuity Systems
- **Developing and applying methods that exploited the knowledge base to analyze experimental data**
  - Network based algorithms, such as PARIS
  - Tools for working with sets (categories)
- **Ran targeted queries against collected knowledge to supply scientists with answers to specific questions**
  - What is known about the cell lines we use?
  - What are transcription factors and targets in pathways of interest?
  - What molecular processes are known to be disease specific?

Effective, but expensive! Let's not repeatedly build the infrastructure for this

Pilot Project: Build a knowledge base - the Neurocommons

- Purpose: develop methods, share how we did it. Make it useful.
- Starting points:
  - Existing information sources (e.g. OBO, Genomic databases, text mining of publications)
  - Representation: OBO Foundry
  - Technology: Semantic Web (RDF/OWL)
- Resulting resource is a “demonstration” (we need this to become a ubiquitous approach. There’s too much for one place to do!)
- Novelty: precise queries across diverse sources

There are a lot of high quality public databases, many with references to the scientific literature



Deep integration
A simple ‘target discovery’ question

Signal transduction pathways are considered to be rich in “druggable” targets - proteins that might respond to chemical therapy

CA1 Pyramidal Neurons are known to be particularly damaged in Alzheimer’s disease.

Casting a wide net, can we find candidate genes known to be involved in signal transduction and active in Pyramidal Neurons?

Mesh: Pyramidal Cell

↓

Pubmed: Journal Articles *(5000)*

↓

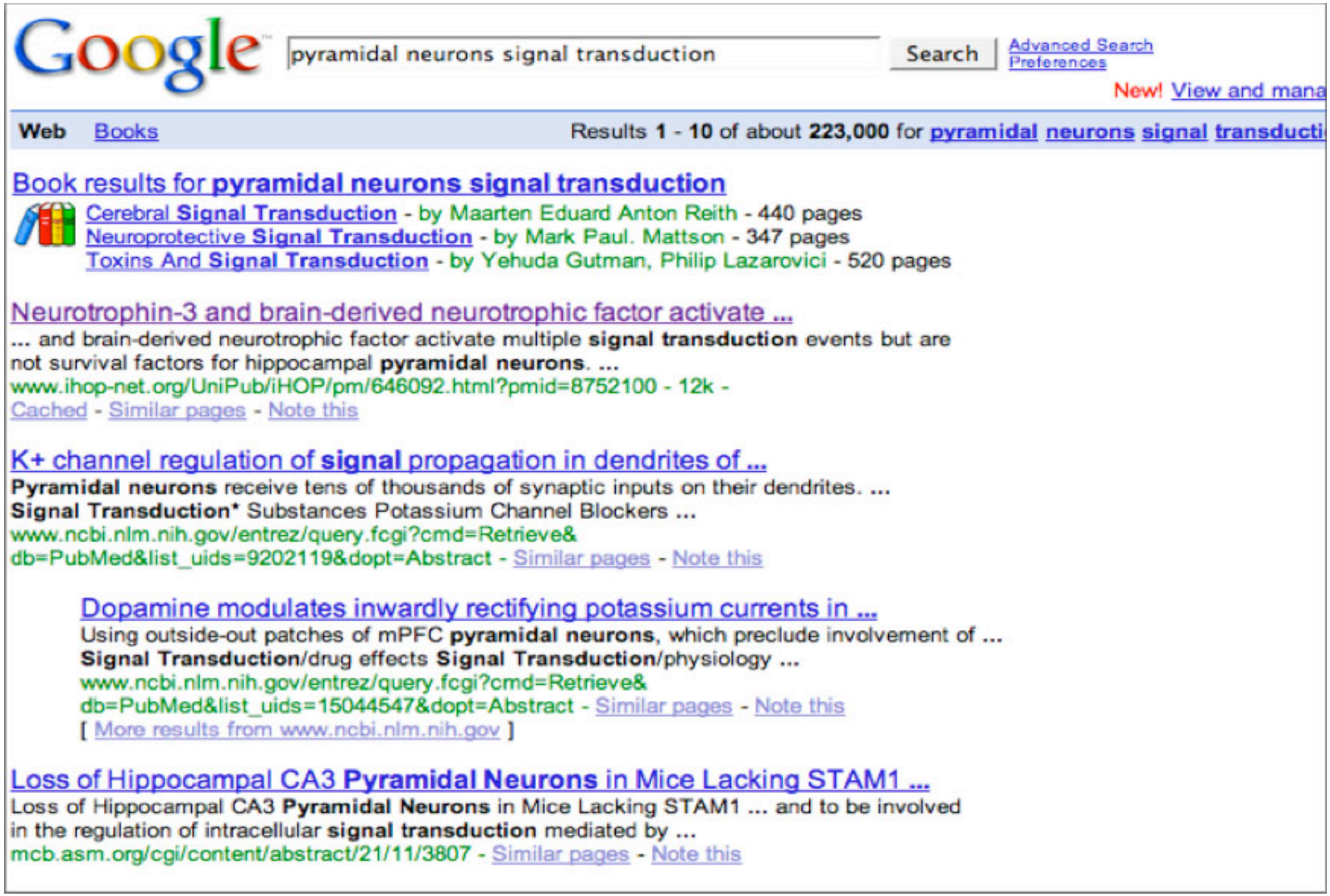
Entrez Gene: Genes

↓

GO: Signal Transduction

All kinds of such processes and parts of them (230)

Google: 223,000 results



Results: genes, processes

DRD1, 1812	adenylate cyclase activation
ADRB2, 154	adenylate cyclase activation
ADRB2, 154	arrestin mediated desensitization of G-protein coupled receptor protein signaling pathway
DRD1IP, 50632	dopamine receptor signaling pathway
DRD1, 1812	dopamine receptor, adenylate cyclase activating pathway
DRD2, 1813	dopamine receptor, adenylate cyclase inhibiting pathway
GRM7, 2917	G-protein coupled receptor protein signaling pathway
GNG3, 2785	G-protein coupled receptor protein signaling pathway
GNG12, 55970	G-protein coupled receptor protein signaling pathway
DRD2, 1813	G-protein coupled receptor protein signaling pathway
ADRB2, 154	G-protein coupled receptor protein signaling pathway
CALM3, 808	G-protein coupled receptor protein signaling pathway
HTR2A, 3356	G-protein coupled receptor protein signaling pathway
DRD1, 1812	G-protein signaling, coupled to cyclic nucleotide second messenger
SSTR5, 6755	G-protein signaling, coupled to cyclic nucleotide second messenger
MTNR1A, 4543	G-protein signaling, coupled to cyclic nucleotide second messenger
CNR2, 1269	G-protein signaling, coupled to cyclic nucleotide second messenger
HTR6, 3362	G-protein signaling, coupled to cyclic nucleotide second messenger
GRIK2, 2898	glutamate signaling pathway
GRIN1, 2902	glutamate signaling pathway
GRIN2A, 2903	glutamate signaling pathway
GRIN2B, 2904	glutamate signaling pathway
ADAM10, 102	integrin-mediated signaling pathway
GRM7, 2917	negative regulation of adenylate cyclase activity
LRP1, 4035	negative regulation of Wnt receptor signaling pathway
ADAM10, 102	Notch receptor processing
ASCL1, 429	Notch signaling pathway
HTR2A, 3356	serotonin receptor signaling pathway
ADRB2, 154	transmembrane receptor protein tyrosine kinase activation (dimerization)
PTPRG, 5793	ransmembrane receptor protein tyrosine kinase signaling pathway
EPHA4, 2043	transmembrane receptor protein tyrosine kinase signaling pathway
NRTN, 4902	transmembrane receptor protein tyrosine kinase signaling pathway
CTNND1, 1500	Wnt receptor signaling pathway

Many of the genes are indeed related to Alzheimer's Disease through gamma secretase (presenilin) activity

What plasmids might be used to study neurodegeneration

Sqarql Query Insert prefix | Insert query | Tools & Snippets

```
prefix owl: <http://www.w3.org/2002/07/owl#>
prefix sc: <http://purl.org/science/owl/sciencecommons/>
prefix foaf: <http://xmlns.com/foaf/0.1/>

select distinct ?submeshname ?title ?plasmidname ?catalogpage ?generecname
  from <http://purl.org/commons/hcls/20070416>
  from <http://purl.org/commons/hcls/20070416/classrelations>
where
{
  # Alzheimer
  ?submesh skos:broader <http://purl.org/commons/record/mesh/D003704>.

  ?pmrec foaf:primaryTopic ?article .
  graph <http://purl.org/commons/hcls/pubmesh>
  { ?pmrec sc:has-as-major-mesh ?submesh.
    ?article dc:title ?title.
  }

  graph <http://purl.org/commons/hcls/20070416>
  { ?plasmid rdfs:label ?plasmidname ;
       sc:is_described_in ?article ;
       sc:carries_DNA_described_by ?generec .
    ?catalogpage foaf:primaryTopic ?plasmid.
    ?submesh skos:prefLabel ?submeshname.
    # ?mesh skos:prefLabel ?meshname.
  }

  graph <http://purl.org/commons/hcls/gene>
  { ?generec rdfs:label ?generecname . }
}
```

Output format: Table | Max Rows: 50 | Execution plan

Run Query | Reset | Clear | Load | Store

result | request | response

submeshname	title	plasmidname	catalogpage	generecname
Alzheimer Disease - metabolism	Neprilysin regulates amyloid Beta peptide levels.	pCSC-SP-PW-Nep (aka: pBOB-NEP)	http://www.addgene.org/pgvec1?f=c&attag=b&cmd=findpl&identifier=12338	Entrez Gene record for mouse Mme, 17380
Alzheimer Disease - metabolism	Neprilysin regulates amyloid Beta peptide levels.	pCSC-SP-PW-NepX (aka: pBOB-NEPX)	http://www.addgene.org/pgvec1?f=c&attag=b&cmd=findpl&identifier=12340	Entrez Gene record for mouse Mme, 17380
Huntington Disease - metabolism	Inaugural Article: A linear lattice model for polyglutamine in CAG-expansion diseases.	pET32a-HD16Q	http://www.addgene.org/pgvec1?f=c&attag=b&cmd=findpl&identifier=11487	Entrez Gene record for human HD, 3064
Huntington Disease - metabolism	Inaugural Article: A linear lattice model for polyglutamine in CAG-expansion diseases.	pET32a-HD25Q	http://www.addgene.org/pgvec1?f=c&attag=b&cmd=findpl&identifier=11508	Entrez Gene record for human HD, 3064

Some questions you care about answering

- For what neurological disorders are cell lines available?
- For Parkinsons disease, what tissue and cell lines are available?
- Give me information on the receptors and channels expressed in cortical neurons
- What chemical agents can be used visualizing the nervous system?

A question I was asked

- Create a system that will let us prioritize an expected 2000 siRNA hits according to whether there is chemical matter for studying them, e.g. validated antibodies, since we can only follow up on 600.

We know how to use Semantic Web technology to answer these kinds of questions

(but there is no free lunch)

Neurocommons technological approach

- From OBO Foundry: Carefully model biology to enable integration of data sources. “Audit trail to reality”
- From Web: Assign all biological entities URIs (lots already provided by OBO). Be on the Semantic Web.
- Translate to RDF and managed in a triple store or provide technology that enables SPARQL query against traditional data stores.
- From OWL: Add triples inferred by reasoner to increase expressiveness of queries with even simple query engine
- From software engineering: Provide data via SPARQL first (API). Build tools on top of that.
- From open source: Make it all completely open and reproducible. Encourage mirrors (2 already)

The Semantic Web

- A standard computable representation for making statements: OWL
- Inference engines that operate on those statements
- A standard query language for programs to ask questions: SPARQL
- All embedded in the Web

The simplest system that might integrate and access knowledge at global scales.

The only viable technology proposed at this scale

Every thing has a web name (URI)

The article: ***Textpresso: an ontology-based information retrieval and extraction system for biological literature***.

The resource was identified by the following URI:
http://purl.org/science/article/pmid/15383839

This URI identifies an article. Access to the article is probably possible, but is not yet implemented via this URI.

Related resources:

- http://purl.org/commons/html/pmid/15383839: HTML representation of PubMed record 15383839, which describes this article
- http://purl.org/commons/record/pmid/15383839: PubMed record 15383839, without commitment as to representation

It's a URL, but more so.

Benefits of Semantic Web technologies for life sciences

- Fusing of data across scientific disciplines is easier due to flexible representation language
  - "connect the NIF to caBIG"
- Single, standardized, web accessible query language
- Querying of data at different levels of granularity and specificity easier because hierarchical representation built into the language
- Emphasis on access, provenance of data
- Ability to perform inference within and across data sets, e.g. "nonsense detection"
- Long term maintenance and hosting through open source community. The Linux model.

Ontology lessons: Three levels of representing scientific knowledge

- *Record level*: Represent database records. Inconsistent if two sources disagree about contents of a field.
- *Statement level*: Represent what researchers say. Inconsistent if two people disagree about what a paper said
- *Domain level*: OBO Foundry approach. Represent your best understanding of consensus. Inconsistent if facts contradict.
- We need all three (but make clear which is which)

Don't inadvertently mix levels

Integration over diverse resources generates challenges for ontology structuring

- Ligand
- Neurotransmitter
- Hormone
- Peptide

Distinguish between thing and function

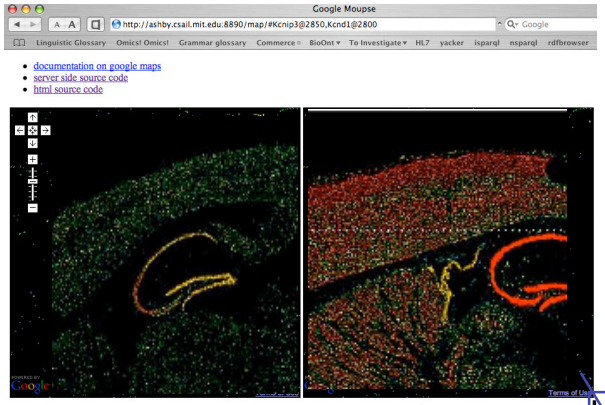
Case study: NeuronDB

- Started with local ontology in EAV. Problem: How to link with everything else
- Some initial issues: No links to evidence, “receptors” versus proteins with receptor activity (like Gene Ontology Association)
- Process, iterate many times, fixing OWL, GO understanding/conformance, augmenting what was in ontology.
- Ends with something that links with GO Function. Accepted process for how to move both NeuronDB and GO forward.
- Interesting: What to do about inconsistency?

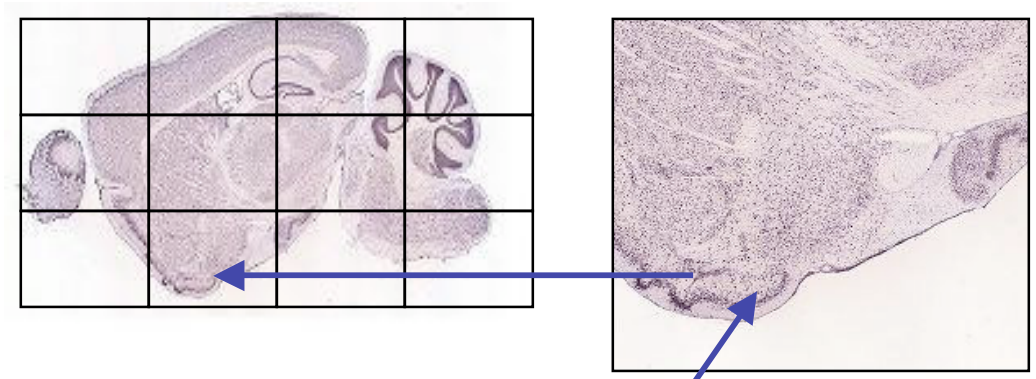
What happens when data is discoverable, queryable, and accessible on the open web?

http://hcls1.csail.mit.edu/map/#Kcnip3@2850,Kcnd1@2800

Javascript



Allen Brain Institute Servers



http://www.brainmap.org://….0205032816_B.aff/TileGroup3/1-0-1.jpg

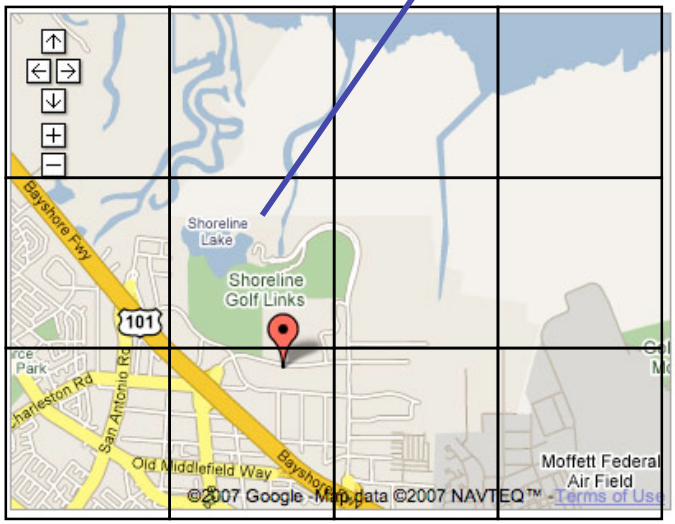
SPARQL
AJAX

Query

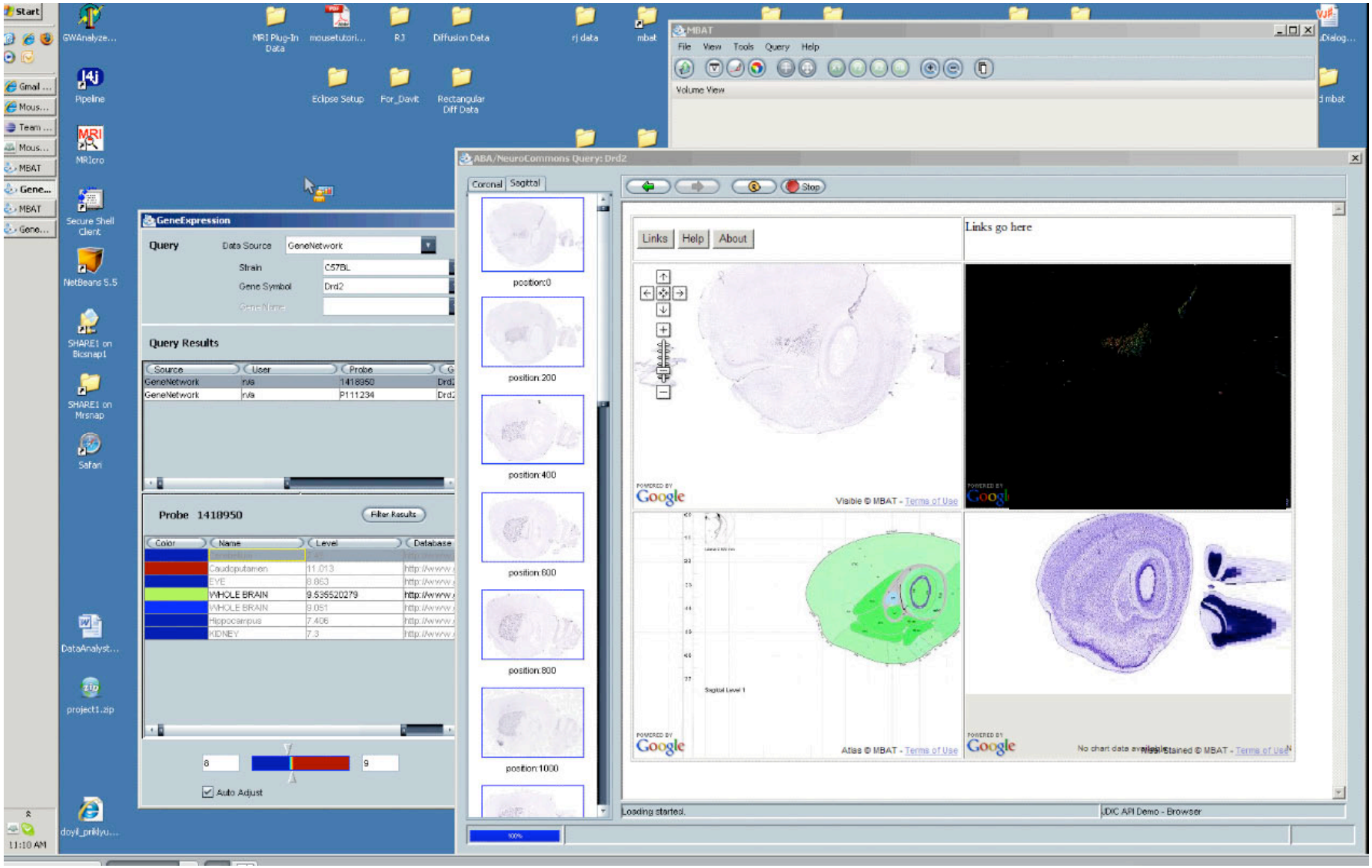
URL

Neurocommons Servers

Google
Maps
API

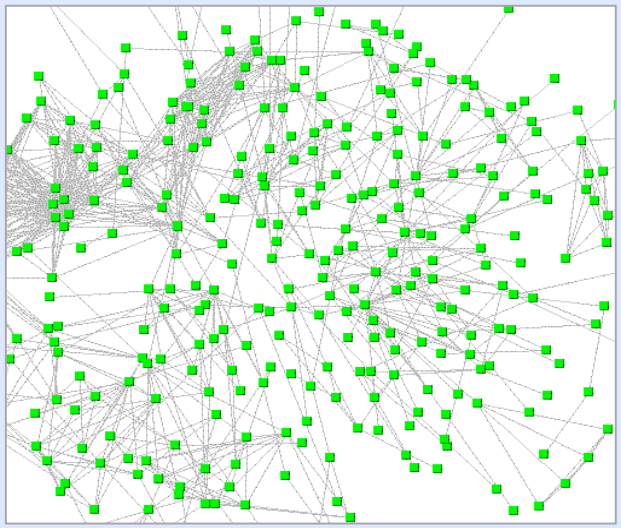


BIRN can “view source”, use our code in MBAT, just like people learning by using others’ html



Activity center analysis

Full Interaction Network

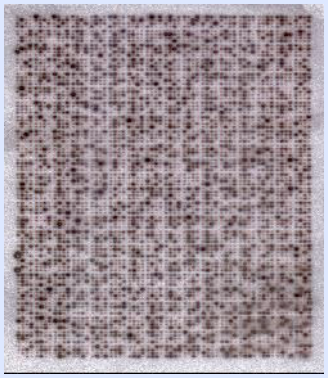


Functional Interactions involving Gene Products

- Binds
- Phosphorylates
- Regulates
- Cleaves…

+

Data, defining *activity*

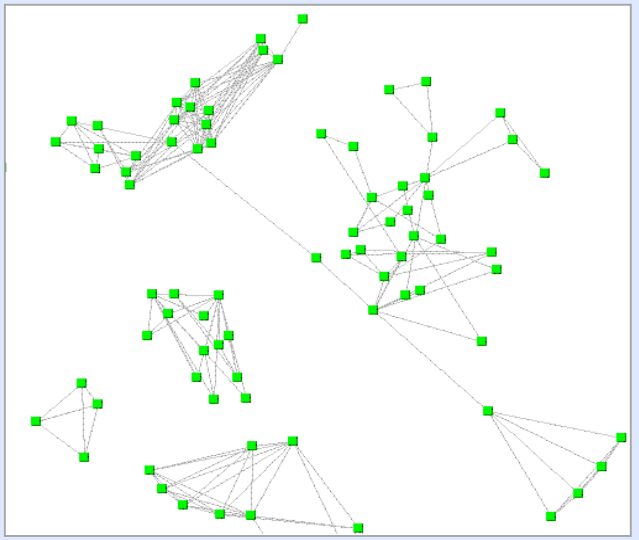


Activity

- Compound vs. Normal
- Knockout vs. Wild Type
- Responders vs. Non-responders

=

***Active* Sub-network**



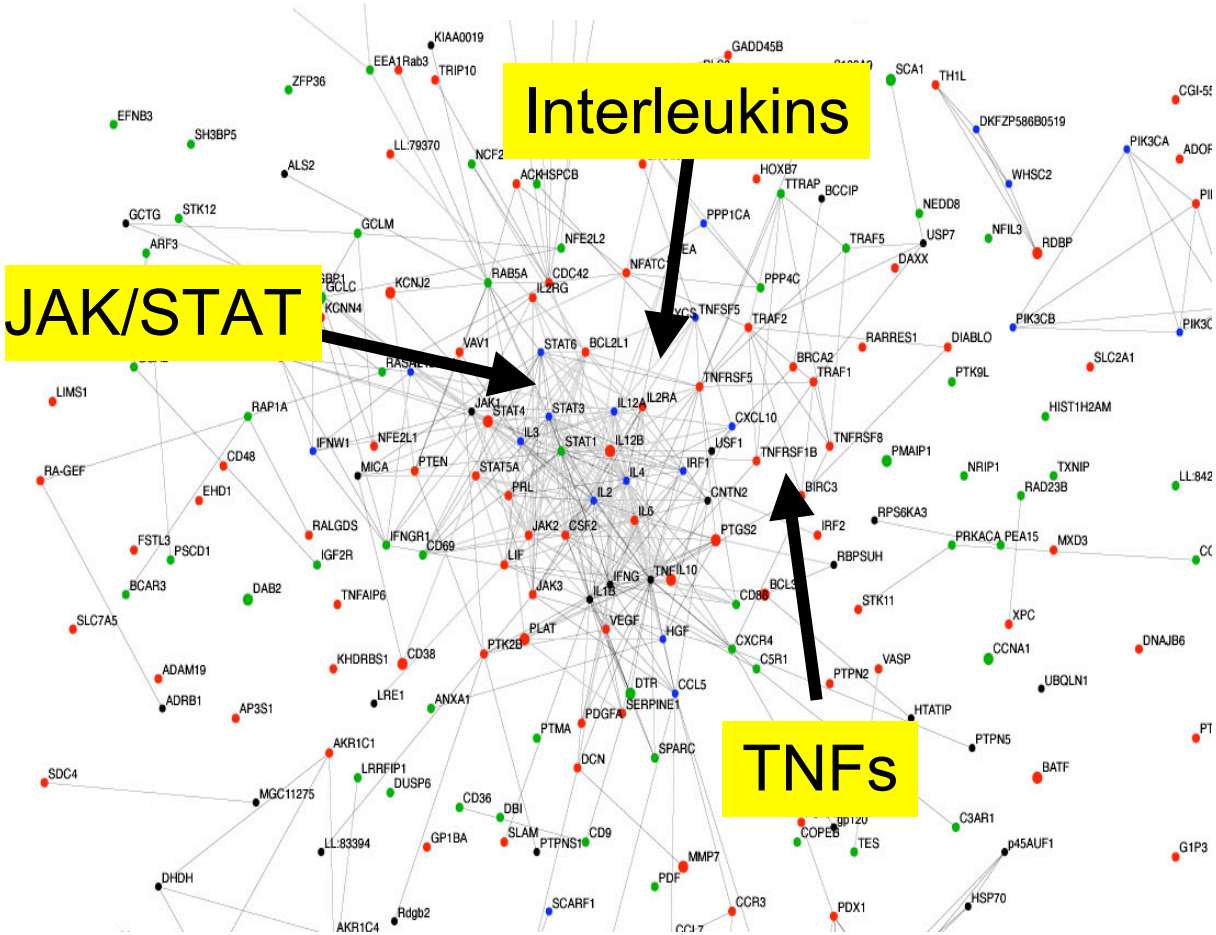
Hints on the Cellular Processes

- Perturbed by a compound
- Downstream of a target
- Involved in drug resistance

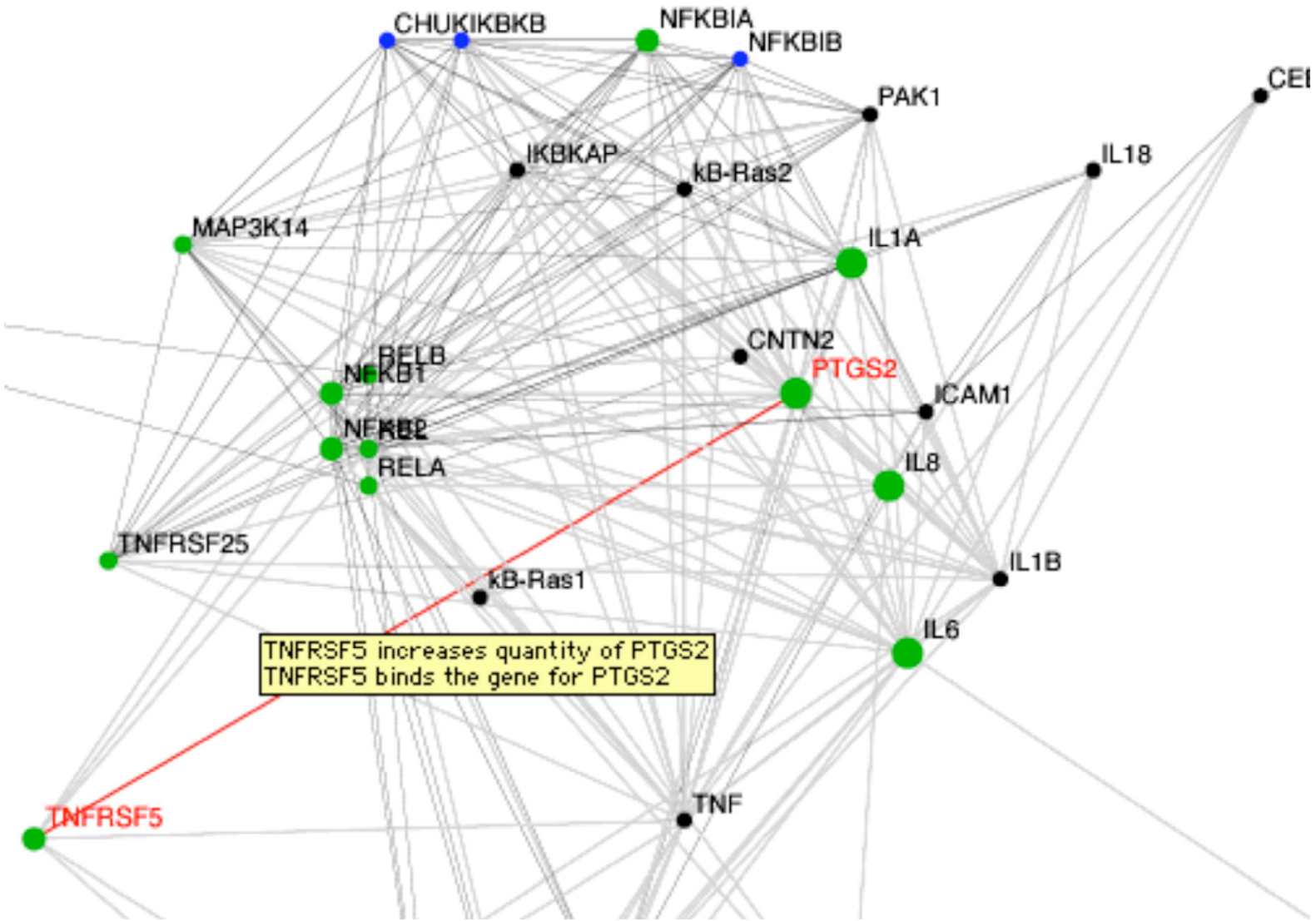
Activity center example: Effects of IKK2 inhibitor

- Primary question: does IKK2 inhibitor pre-treatment reduce response to LPS stimulation?
- Stimulating 14-day-old monocyte-derived dendritic cells with treated with LPS alone with or an IKK2 inhibitor.
- Transcript profiling (Affymetrix U133) timepoints are 4.5 hours and 24 hours post LPS treatment.

IKK2 Inhibitor + LPS vs. LPS, 4.5h: Inflammatory pathways down-regulated



LPS



Code available under BSD open source license from Science Commons
Releasing web tool for selecting GEO datasets to analyze using the Neurocommons

What's on our agenda

- Connect more knowledge bases - cells, anatomy, physiology, behavior, protocols, reagents
- Beyond simple interaction: More precise representations of mechanism to be able to query and exploit computationally
- Built in a open, scalable, scientifically credible way, to encourage sustained contribution, and to take advantage of "web effects"
- Capture relationships generated by text mining efforts such as Textspresso, Powerset, gopubmed, etc.
- Make the Neurocommons and our approach useful within the NIF (and vice versa)

How do we get there?

- Interoperation is paramount, but modeling is hard: Work with the OBO Foundry
- Build a skilled community by encouraging collaboration and apprenticeship at Science Commons
- Use (open!) Semantic Web Technologies to enable web effects and global scope
- Support and nurture a growing and vigorous community (BIRN, caBIG, SWAN, OBI) all of whom build on the rest and enable others to build more
- Work to advance key technologies and infrastructure - text mining, structured abstracts, query, reasoning.

Selected Links

- http://sw.neurocommons.org/
- http://esw.w3.org/topic/HCLS/Banff2007Demo
- http://hcls1.csail.mit.edu:8890/nsparql/
- http://hcls1.csail.mit.edu:8890/map/#Kcnip3@2850,Kcnd1@2800
- http://obi.sourceforge.net/
- http://neuroweb.med.yale.edu/senselab/

Acknowledgements

- W3C Semantic Web Health Care and Life Sciences Interest Group
- Senselab: Kei Cheung, Matthias Samwald
- OBO/OBO Foundry, esp. Barry Smith, Chris Mungall
- BIRN: Bill Bug, Maryanne Martone
- Jonathan Rees, Joel Pradines
- Millennium Pharmaceuticals